Journal of Experimental Psychology
1973, Vol. 101, No. 2, 360–366

ENCODING AND RECOGNITION MEMORY FOR NATURALISTIC SOUNDS [1]

GORDON H. BOWER [2] AND KEITH HOLYOAK

Stanford University

The question examined was whether a naturalistic sound is remembered chiefly in terms of the "object interpretation" *S* assigns to it (e.g., sonar pings, crickets chirping). If so, and if the sound is reinterpreted in a different way during testing, then *S* should not remember having heard it in an earlier set. Subjects listened to a set of ambiguous sounds and received a later recognition memory test with *old* and *new* stimuli. Some *S*s labeled the sounds themselves whereas others had the sounds labeled for them during study and testing. Recognition memory was much higher for *old* sounds given their old labels during testing than for those receiving new (though plausible) labels, with equal effects for *S*-labeled and *E*-labeled stimuli. Moreover, false alarms to *new* sounds were partly determined by their eliciting the labels of *old* stimuli. It is concluded that recognition memory largely reflects the similarity of the referential interpretations that *S* provides for the study and test sounds.

This study investigates the influence on recognition memory of the perceptual organization of the learning material both at the time it is studied and when it is tested. Most of the earlier work on this topic has used visual stimuli (e.g., Leeper, 1970) or verbal materials (e.g., Light & Carter-Sobell, 1970). The present work extends these investigations to "naturalistic sounds," a relatively unexplored stimulus domain for psychologists.

It is assumed that when *S* hears a naturalistic sound (e.g., a carpet sweeper, crickets chirping), he tries to interpret the referent of that sound, perhaps by comparing it with a set of prototypes in memory. When the test is sufficiently close to a prototype sound, we will say that *S* assimilates the input to that prototype. In a recognition memory experiment in which *S* hears many different naturalistic sounds to remember, it is assumed that he establishes an association between each "sound-as-interpreted" (the referent) and a context marker. This association essentially encodes the proposition, "In the study list I heard a sound like that made by an *X*" (see Anderson & Bower, 1972). If this association is formed and if the later presentation of the same sound in a test list is also reinterpreted as "an *X*," then *S* will remember that he heard this sound in the study list. In this theory, then, recognition memory requires that the sound be encoded or interpreted in essentially the same way during studying and testing.

Unambiguous stimuli are of little value for testing this hypothesis, since they are reliably identified in an invariant way. More interesting possibilities are provided either by ambiguous stimuli, which can be interpreted in several ways, or by vague patterns, which are hard to organize (which require much searching to find any interpretation, possibly with many failures). In the visual domain, examples of ambiguous pictures are the "wife vs. mother-in-law" picture or the "duck–rabbit" picture, whereas vague, hard-to-organize patterns are exemplified by the Mooney pictures with deleted contours (Mooney, 1957) or incomplete object figures (Leeper, 1935). Ambiguous stimuli are useful for testing the recognition hypothesis insofar as contextual factors can influence how they are organized and interpreted (Leeper, 1935; Light & Carter-Sobell, 1970). Hard-to-organize figures are similarly useful insofar as an *S* either

[1] This research was supported by Grant MH-13950-06 to the first author from the National Institute of Mental Health. The second author was supported by a Stanford University fellowship.

[2] Requests for reprints should be sent to Gordon H. Bower, Department of Psychology, Stanford University, Stanford, California 94305.

succeeds or fails in seeing the complete figure, and his later recognition memory will be poor if he fails to see something or if during testing he sees something he failed to see before (Wiseman & Neisser, 1971). In order to manipulate stimulus encoding, the following experiment on sound recognition used as critical stimuli sounds that were prerated by listeners as ambiguous, with several plausible but different interpretations.

The question of interest is how recognition memory depends on the similarity of encoding of the sound during study and testing. We have the option of either letting *S* determine how to interpret an ambiguous sound, or having *E* suggest a particular interpretation as *S* listens to the sound. Since the encoding configurations of interest may occur in either case, both conditions were run in this experiment. For one group *E* supplied interpretations for each sound *S* heard during study; the studied sounds were then divided into equal subgroups, and were later tested by being played along with the same label, a different (but equally plausible) label, or no label (*S* was asked to either remember *E*'s earlier label or to supply his own for the sound). For the second group, while *E* labeled a few of the study sounds, *S* tried to interpret and label the majority of them; at the time of testing, these *S*s tried to label each sound and to judge whether that exact sound had occurred in the study list. In either case, recognition memory for the sound should be poorer either if *S* achieves no interpretation of it (at input or at test) or if the interpretation he achieves during testing differs substantially from that achieved during study.

Method

Design. Each *S* listened to a study tape of sounds (stimuli), then returned a week later for a recognition memory test with a second tape composed of half *old* and half *new* stimuli. The design of the experiment is shown in Table 1, which depicts the composition of the study and test lists for the 2 groups. The study list of 80 sounds was comprised of 60 critical sounds preselected for their ambiguity, and 20 easily identified "fillers" included to increase task difficulty

TABLE 1

Schematic Composition of Study and Test Lists for Groups EL and SL

Study list	Test list
Group EL	
60 EL	20 *old* + same EL 20 *old* + SL 20 *old* + different EL
+20 EL fillers	40 *new* + EL 20 *new* + SL
Group SL	
40 SL 20 EL	60 *old* + SL
+20 fillers (10 SL, 10 EL)	60 *new* + SL

Note. The terms EL and SL denote *E*-labeled and *S*-labeled stimuli, respectively.

but which were not tested later. For Group EL (*E* labeled), during the study trial each sound was preceded by an appropriate *E* label. The 60 critical stimuli were then divided into 3 sublists of 20, to be presented randomized on the test list either with the same label, a different (though also appropriate) label, or no label. These 60 *old* sounds were mixed in with 60 *new* test sounds, 40 of which were labeled by *E* while 20 were not. Whenever *E* did not supply a label, *S* was asked either to recall or to think up a label for the sound (denoted *S* labeled or SL henceforth). For the second condition, Group SL, *S* labeling of the sounds was emphasized; $\frac{2}{3}$ of the study-list stimuli and all of the test-list stimuli were to be labeled by *S*. For each test stimulus, *S* was first to decide how to label it, and then decide whether or not he had heard that sound in the study list. He judged *old* or *new*, and also rated the confidence of his judgment on a 3-point scale.

Materials. The study list contained 80 sounds (60 critical ones, 20 fillers), and the test list contained the 60 *old* critical sounds mixed with 60 *new* sounds. The 140 sounds were selected from several albums of "sound effects" records produced commercially by Jac Holzman and Audio Fidelity, Inc. We recorded the first 12 ± 4 sec. of each sound strip on a tape recorder for the actual presentation. The 60 critical stimuli were preselected by a panel of 3 listeners to be ambiguous; that is, these were sounds for which at least 2 different interpretations were plausible and were commonly supplied (one of these was invariably the label for the sound supplied on the record jacket). Three example pairs were sonar pings vs. jungle insects, heartbeat vs. bouncing rubber ball, and soldiers clumping down stairs vs. horses trotting through the streets.

TABLE 2

RELATIVE OLDNESS RATINGS (R_i) BY GROUP EL FOR *OLD* AND *NEW* TEST SOUNDS DEPENDING UPON THEIR LABELING BY *E* DURING THE TEST

Stimulus	Label	R_i score	Corrected
Old	Same L	.80	.67
	No L	.59	.29
	Different L	.47	.19
New	L	.34	—
	No L	.43	—

Note. The corrected score is the difference in scores for *old* minus *new* stimuli divided by 1 minus the score for *new* stimuli (EL = *E* labeled; L = label.)

For Group EL, only one study list was constructed, using 1 of the 2 labels selected randomly. Three different test lists were then composed, each used with 4 *S*s of Group EL; the test lists differed according to which critical stimuli retained the same label as during study, were changed to a different label, or were tested with no label. Across *S*s in Group EL, each critical sound was tested equally often in the 3 conditions of same label, different label, or no label.

For Group SL, 3 study lists were constructed differing according to which subset of 20 critical sounds were labeled by *E*. Five *S*s learned each study list. For Group SL there was only one test list since neither *old* nor *new* stimuli were labeled. As an incidental task during the study trial, *S*s in Group SL also rated the "fittingness" or "appropriateness" of the label given (by *E* or *S*) to each sound. This rating was on a 5-point scale from *very poor* to *very good* fit between label and sound.

Procedure. The *S*s listened to the tape-recorded sounds in small groups of 2–6. Each sound stretch was 12 ± 4 sec., with a 5-sec. pause between the sounds. If *E* supplied a label (1–5-word phrase), it was given in the interval preceding the sound. The rate of stimulus presentations was the same during testing as studying. In case no *E* label was given for Group EL, *E* said "Your guess!" just before the sound came on, indicating that *S* was to write down on a numbered answer sheet his interpretation of the sound. Since *S*s in Group SL were to label *all* sounds on their test list, they were not told "Your guess" before each test sound, but simply wrote their interpretations. The retention test sheet contained 120 blanks (for labels, if needed) alongside the 6-point recognition rating scale ranging from 1 = *sure old* though 6 = *sure new*. It was carefully explained to *S*s that they should make their recognition memory judgments with respect to the sounds per se, since some of the labels would be changed. No fittingness judgments were required during recognition testing.

Subjects. The *S*s were 27 Stanford undergraduates; 18 were paid $3.00 for their participation whereas 9 were fulfilling a service requirement for their introductory psychology course. There were 12 *S*s in Group EL and 15 in Group SL, with approximately equal numbers of males and females in the 2 groups.

Scoring protocols. Several measures were derived from the recognition protocols. One measure for SL test stimuli is the proportion of cases when *S* recalled during testing the same label as the one given to that sound (by *E* or *S*) during the study trial. A lenient criterion of synonymity was used for classifying 2 labels as substantially the same in meaning. The recognition responses were considered first as dichotomous observations, yielding proportion of *old* judgments for each class of stimuli. A second measure used the 6-point confidence rating, rescaling the mean score so that it is interpretable as *S*'s average rating of the likelihood that test items of a given class are *old*. If C_i is the average confidence rating for items of Type i (where 1 = *sure old*), then the relative oldness rating, R_i, was defined as $(6 - C_i)/(6 - 1)$. This R_i index varies from 0 to 1 as the average judgment ranges from *sure new* to *sure old*. The R_i index can be viewed as an estimate of a probability (see Green & Swets, 1966) and corrected for false alarms.

RESULTS

Condition EL. The exposition is simplified if the results from the 2 groups are discussed separately. The main results for Group EL are shown in Table 2 which gives the average relative oldness ratings for *old* and *new* test stimuli depending on the labeling of the sound during testing. For *old* stimuli, recognition is best when they are heard in the context of the same *E*-supplied label and poorest when *E*'s test label differs from that supplied during the study trial. An overall analysis of variance on the transformed (arc sine) relative oldness scores for the 3 classes of *old* stimuli yielded a significant effect, $F\ (2,\ 22) = 32.8,\ p < .001$. Newman-Keuls comparisons (for correlated measures) revealed that all differences among the R_i measures to *old* stimuli for the 3 conditions were significant ($p < .01$). Essentially similar results and significance levels were obtained with dichotomous scoring of *old* vs. *new* recognition responses. The corrected relative oldness scores in the final column of Table 2 are $(R_o - R_n)/(1 - R_n)$, where R_o and R_n are the mean relative oldness ratings of items in particular *old* and *new* conditions (with or without labels). This is like the standard "correction for false alarm rates"

in high-threshold detection theory (see Green & Swets, 1966). Note that R_n differs depending on the presence or absence of a label for the *new* stimulus. (This difference in R_n falls short of statistical significance.) Another analysis of variance on the corrected oldness ratings for the 3 classes of *old* stimuli produced outcomes similar to those of the former analyses with uncorrected scores.

Each *S* in Group EL received 20 unlabeled test stimuli and was asked to supply (or remember) a label for each. We may categorize *S*'s label as being either substantially the same as or different from the label *E* originally gave for that sound during the study trial. Only 30% of the test stimulus labels were categorized as the same as their mate given by *E* during the study trial (were "remembered"). However, this sameness of encoding strongly influenced recognition memory for SL sounds. *Old* sounds which *S* labeled the same as during training were rated as *old* 84% of the time, whereas sounds for which *S* provided a substantially different label during testing were rated as *old* only 45% of the time. These proportions differ reliably, χ^2 (1) = 8.41, $p < .01$.

When we say that *S* gave a test stimulus a label different from that supplied for it by *E* during the study trial, we do not exclude the possibility that *S* labeled the test sound by a label which *E* had given to some other *old* sound during training. If *S*'s interpretations are a major determinant of his recognition memory decisions, then we may expect any sound to which *S* assigns an *old* label to be called *old* with a high probability, regardless of the history of the test stimulus. The clearest demonstration of this control by reactivation of a prior label is provided by the *S*-labeled *new* stimuli. We examined the 240 cases of labels supplied by EL *S*s to their new unlabeled test stimuli. In 46 cases *S*s gave to a *new* stimulus an *old E* label from the study list; of these cases, 33, or 72%, were judged to be *old* sounds. This false alarm rate contrasts with that for the remaining *new* stimuli which received *new* labels; these were called *old* only 27% of the time. Clearly, then, *S*'s sound recognition is being partly mediated through his interpretation of the sound, and this correlation is as strong for false alarms to *new* stimuli as for correct recognitions to *old* stimuli.

TABLE 3

RELATIVE OLDNESS RATINGS (R_i) FOR CATEGORIES OF TEST STIMULI (Group SL), AND PERCENTAGE OF RESPONSES OF VARIOUS CATEGORIES

S label	Percentage of item type	R_i score	Corrected
	Old:EL		
Same	35	.74	.57
Different	65	.53	.23
	Old:SL		
Same	41	.84	.74
Different	50	.54	.25
Not labeled on study	9	.46	.11
	New		
—	—	.39	—

Note. On the study trial, Old:EL items were labeled by *E*; Old:SL items were left for *S* to label. Either the same or a different label was generated by *S* on the test trial. Data are tabulated separately for sounds which *S* was unable to label during study.

Condition SL. The results for Group SL are summarized in Table 3. Test items have been categorized according to whether they are *old* or *new*, and *old* stimuli are further divided according to whether they were *E* labeled or *S* labeled during study and whether the *S* label during testing was the same as or different from the study label. The average relative oldness rating is reported for each item category; these ratings corrected for performance to the *new* stimuli, $(R_o - R_n)/(1 - R_n)$, are shown in the right-hand column.

As in Group EL, the probability that S labeled an *old* test sound with the same label it received during study was just slightly over $\frac{1}{3}$. There was no reliable difference in label memory for the *E*-labeled and *S*-labeled stimuli (i.e., 35% vs. 41%). However, in each case, there was a strong influence of the identity of the

study and test labels on recognition memory: Test sounds receiving the same label were recognized at a higher level than test sounds receiving a different label. A 2 × 2 analysis of variance for repeated measures was carried out on the ratings for the 4 conditions defined by the factorial combination of *E* vs. *S* labeling and same vs. different labeling during testing. Same-labeled stimuli were rated reliably higher than differently labeled stimuli, F (1, 14) = 125, $p < .001$. When *S* recalled the same label, he was somewhat more likely to recognize the sound as *old* if he rather than *E* had labeled the sound during study; however, this difference was not statistically reliable. Performance was poorest on that small set of items that *S* attempted but failed to label on the study trial; these are stimuli for which *S* could find no interpretation, and they are not reliably discriminated during testing from new stimuli (i.e., R_i scores of .46 vs. .39). Comparing performance across Groups EL and SL (Tables 2 and 3), the overall levels and patterns are rather similar. What seems important for recognition is not whether *S* or *E* labeled the stimuli during study but whether *S* reinterprets the sound during testing the same way it had been labeled during study.

Recall that *S*s in Group SL rated each of their own (or *E*'s) labels for each study sound on a 5-point fittingness scale. We examined the correlation between judged label fittingness and subsequent recognition by tabulating all judgments of *old* test stimuli in a 5 (Fittingness Ratings) × 6 (Later Recognition Ratings) contingency matrix. The contingency correlation was an insignificant −.05. This low correlation may only reflect a compressed range of "true" fittingness judgments, since all of *E*'s labels and most of *S*'s presumably were chosen so that the labels were fitting and plausible interpretations of the sounds. An independent experiment would be required to vary this fittingness attribute over a larger and possibly more potent range (see the influence of label associativity on shape recognition, Ellis, 1973).

One may ask why *old* but differently labeled sounds were recognized as *old* more often than *new* unlabeled stimuli (see Tables 2 and 3). A first possibility, especially likely for *S*s in Group EL, is that an *old* test sound (or one or more of its distinctive features) is remembered per se and that it elicits the label *E* had associated with it during the study trial. In this case, *S* would recognize the sound as *old* despite the changed *E* label during testing—in fact, *S* may notice the ambiguity of the sound or even reject the changed *E* label as implausible. A second possible explanation stems from the fact that the *old* stimuli were preselected so that they would be ambiguous, whereas the *new* stimulus lures were picked from the remaining pool of relatively unambiguous sounds. Thus, the *new* stimuli were probably somewhat more likely to elicit a confident label, and because that label was new, the sound was likely to be identified as *new* with high confidence. Therefore, this stimulus materials factor could account for part of the difference in recognition between *old* differently labeled and *new* unlabeled stimuli.

A third possibility is that *S* uses a different criterion than does *E* for deciding what are same vs. different labels (interpretations) for the same sound. To illustrate, suppose that during the study trial *S* (or *E*) labeled a particular sound "car," whereas he labeled the corresponding test sound "engine" or "motor" or "vehicle" or "bus" or "trolley" or "motor-driven pump," and so on. Which of these label pairs are the same and which are different? The question is central to our enterprise but is rather difficult to answer. Obviously, the same–different dichotomy results from just one (vague) cut point on an underlying continuum of "referential similarity" of the 2 labels for *S*. In scoring *S*'s labels in Group SL, we were lenient in accepting semantically similar labels as same (less than 5% of the label pairs seemed questionable). In selecting the critical sounds and label pairs for Group EL, we chose only those for which the primary referents of the labels seemed intuitively far apart. Nevertheless, to

check on the plausibility of this label similarity factor, we reexamined our label pairs and selected 20 pairs (of the 60 EL sounds) for which the pair of labels seemed more similar referentially than was the case for the remaining 40 pairs. We shall refer to these as the "close" and the "distant" different sets of label pairs. We then examined recognition ratings for Group EL for these 2 sets of sounds, when no label was given at testing compared with when the different label was given. The results were that for the set of 20 close labels, probability of correctly responding *old* to unlabeled tests was .69 whereas to differently labeled tests it was .56. These do not differ reliably. On the other hand, for the set of distant labels, the corresponding probability for unlabeled tests was .57 vs. .39 for differently labeled tests. These latter percentages do differ reliably, χ^2 (1) = 9.37, $p < .01$. Thus, even in terms of this relatively crude classification of different labels for the same sound as close or distant from one another in meaning, one finds a greater loss in recognition (with respect to the corresponding control items) for test items whose label is referentially more distant from the label studied.

Discussion

The present experiment fits into a long tradition of research on the influence of verbal encoding upon memory for nonverbal materials. Most of that research has used visual shapes (see Ellis, 1973, for a recent review). Naturalistic sounds would appear to provide a rich domain of stimulus materials for use in psychological experiments. Our data showed that while the label or "interpretation" of the test sound was very influential in *S*'s recognition decision, it did not totally dominate that judgment. There was still a residual effect due to the test sound being *old* or *new*, irrespective of how it was labeled. It would thus appear that *S*'s memory trace is a composite system comprised of some crude physical description of the sound (e.g., a salient feature) associated with a meaningful label, this entire complex associated with a list marker of the form, "I heard this *X* in the study list." The recognition rating given to a test sound is presumably higher the greater the degree to which its internal representation or encoding overlaps with the composite memory trace of its corresponding target from the study list. The present results suggest that the label (or "referential situation") carries relatively greater weight than do the sensory parameters in determining the recognition ratings. This weighting seems introspectively correct: We typically remember a nonmusical sound sequence by recalling a scenario of visual images of the objects making those sounds. If one alters the referential imagery evoked, one correspondingly alters the identification of the sound.

An elementaristic analysis of recognition memory would suppose that recognition ratings depend systematically upon the degree of change in some sensory parameters of the sound pattern (e.g., loudness would probably be an inconsequential variable) and the change in the referential meaning of the label given to the sound. An *E* label not only serves to interpret the sound but also would be expected to focus attention upon different distinctive (acoustic) features of the sound train. Thus, even the basic "sensory description" of the sound would vary depending on the label.

However, one should not overemphasize this referential biasing of sensory descriptions. After all, during a lifetime we obviously learn reasonably accurate sensory (acoustic) descriptions of the common sounds things make, since it is with these acoustic descriptions that we classify a sound as a new exemplar of a category (e.g., a barking dog, a car backfiring). However, once these categories are established, later descriptions of sounds (e.g., as in our recognition memory studies) become assimilated to the categories, and the category prototype seems to replace the unique set of sensory parameters characterizing each exemplar. These categorical judgments, in turn, seem to exert a potent influence over recognition memory of the sort studied here. This categorical determination of recognition memory has an analogue in the visual modality with ambiguous figures: If *S* is biased to interpret the duck–rabbit figure as a duck, he cannot later retrieve from memory the picture's exact appearance to recategorize it as a rabbit.

There still remain many puzzles concerning the interplay between sensory situations, referential labels, and recognition memory. First, it is well known that the same stimulus event can be described with degrees of detail

varying all the way from the specification "the Vienna Boy's Choir singing Handel's *Messiah*" through "a choir singing music" to "human voices" or even "sounds" as the most general category. We may suppose that a person identifies (or describes) an event with the degree of specificity most frequent for his repertoire of discriminative skills (e.g., compare a concertmaster's with an aborigine's identification of a piece of symphonic music). Accordingly, an *S*'s interpretation is not likely to be influenced by providing him during study with a label which identifies a more general and less discriminating category than he can already provide on his own. Similarly, during testing with *E*-labeled stimuli, the more general and abstract the *E* label, the less likely it would be to selectively retrieve a more highly specified memory trace of the appropriate study stimulus. The important issue in recognition memory experiments is whether the test label permits retrieval of traces of a prior interpretation of the stimulus pattern as well as providing for differentiation of an *old* from a *new* set of items. For example, a test EL such as "This is a mechanical sound" is too nonspecific to be helpful when the interpretation *S* must achieve is "carpet sweeper" and what he must reject is "Waring Blendor." Research on this important role of label specificity on recognition memory is yet to be done.

A second, independent issue awaiting research is whether recognition memory for a particular sound is influenced by the organized setting in which it occurs. For example, a Gestalt psychologist would be concerned with the "laws" determining how the sound S_2 becomes "unrecognizable *as* S_2" when it is embedded in the context of other sounds as $S_0S_1S_2S_3S_4$; the context may bring out "emergent properties," either altering the meaningful interpretation of S_2 or altering its distinctiveness so that *S* may no longer be able to segregate S_2 out of the pattern as a unit to be recognized.

This discussion illustrates the research potentialities awaiting exploration with naturalistic sounds. Our preliminary findings established strong communalities with visual form perception: Recognition memory for nonverbal stimuli is enhanced if a label suggests the same perceptual interpretation for it at the time of its study and test, whereas recognition is seriously impaired if a perceptual interpretation is not achieved, or is completely altered to a different referent or meaning. Although there appears to be some memory for naturalistic sounds per se, that effect is small relative to the determination by the perceptual interpretation of the stimuli.

REFERENCES

ANDERSON, J. A., & BOWER, G. H. Recognition and retrieval processes in free recall. *Psychological Review*, 1972, **79**, 97–123.

ELLIS, H. C. Stimulus encoding processes in human learning and memory. In G. H. Bower (Ed.), *The psychology of learning and motivation*. Vol. 7. New York: Academic Press, 1973.

GREEN, D. M., & SWETS, J. A. *Signal detection theory and psychophysics*. New York: Wiley, 1966.

LEEPER, R. W. A study of a neglected portion of the learning field—The development of sensory organization. *Journal of Genetic Psychology*, 1935, **46**, 41–75.

LEEPER, R. W. Cognitive learning theory. Part V. In M. H. Marx (Ed.), *Learning theories*. New York: Macmillan, 1970.

LIGHT, L. L., & CARTER-SOBELL, L. Effects of changed semantic context on recognition memory. *Journal of Verbal Learning and Verbal Behavior*, 1970, **9**, 1–11.

MOONEY, C. M. Age in the development of closure ability in children. *Canadian Journal of Psychology*, 1957, **11**, 219–226.

WISEMAN, S., & NEISSER, U. Perceptual organization as a determinant of visual recognition memory. Paper presented at the meeting of the Eastern Psychological Association, New York, April 1971.

(Received April 17, 1973)